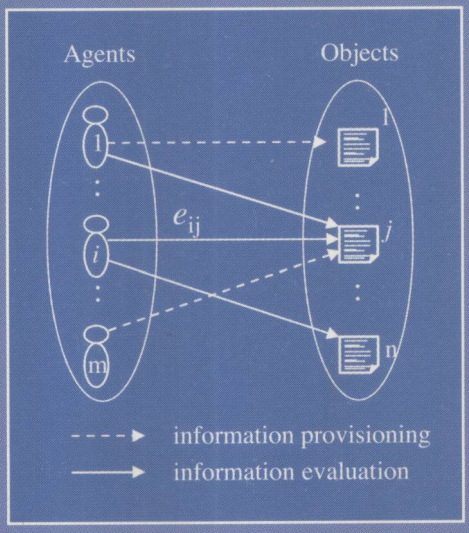
Rino Falcone
Suzanne Barber
Jordi Sabater-Mir
Munindar P. Singh (Eds.)

Hot Topics
LNAI 3577

Trusting Agents for Trusting Electronic Societies

Theory and Applications in HCI and E-Commerce



Springer

Rino Falcone Suzanne Barber
Jordi Sabater-Mir Munindar P. Singh (Eds.)

Trusting Agents for Trusting Electronic Societies

Theory and Applications in HCI and E-Commerce

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Rino Falcone
National Research Council, Institute of Cognitive Science and Technology
Artificial Intelligence Group
Via San Martino della Battaglia 44, 00185 Rome, Italy
E-mail: r.falcone@istc.cnr.it

Suzanne Barber
University of Texas at Austin, Electrical and Computer Engineering
The Laboratory for Intelligent Processes and Systems, Austin, TX 78712, USA
E-mail: barber@cadlips.ece.utexas.edu

Jordi Sabater-Mir
National Research Council, Institute of Cognitive Science and Technology
Agent Based Social Simulation Laboratory
Via San Martino della Battaglia 44, 00185 Rome, Italy
E-mail: jsabater@iiia.csic.es

Munindar P. Singh
North Carolina State University, Department of Computer Science
940 Main Campus Drive, Suite 110, Raleigh, NC 27606, USA
E-mail: singh@ncsu.edu

Library of Congress Control Number: 2005929386

CR Subject Classification (1998): I.2.11, I.2, H.5.3, K.4, C.2.4

ISSN 0302-9743
ISBN-10 3-540-28012-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28012-5 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11532095 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 3577

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Preface

This special issue is the result of two workshops, the 6th and 7th International Workshops on Trust in Agent Societies, respectively held in Melbourne (Australia) on July 14, 2003 and in New York (USA) on July 19, 2004 as part of the Autonomous Agents and Multi-agent Systems 2003 and 2004 conferences (AAMAS 2003 and AAMAS 2004), and organized by Rino Falcone, Suzanne Barber, Larry Korba, and Munindar Singh (AAMAS 2003) and by Rino Falcone, Suzanne Barber, Jordi Sabater-Mir, and Munindar Singh (AAMAS 2004).

The aim of the workshops was to bring together researchers from different fields (artificial intelligence, multi-agent systems, cognitive science, game theory, and the social and organizational sciences) to contribute to a better understanding of trust, reputation and security in agent societies. The workshops' scope included theoretical results as well their applications in human-computer interaction and electronic commerce.

This volume includes a selection of the revised and extended versions of the works presented at the two workshops, incorporating many points that emerged in the discussions, as well as invited papers from expert people in the field. In our view the volume gives a complete coverage of all relevant issues.

We gratefully acknowledge the financial support from the Italian National Research Council — Institute of Cognitive Sciences and Technology, by the European Project MindRACES (from Reactive to Anticipatory Cognitive Embodied Systems; contract number. FP6-511931), and by the Marie Curie Intra-European fellowship contract number MEIF-CT-2003-500573.

May 2005

Rino Falcone
Suzanne Barber
Jordi Sabater-Mir
Munindar Singh

Sponsoring Institutions

Italian National Research Council — Institute of Cognitive Sciences and Technologies

MindRACES (from Reactive to Anticipatory Cognitive Embodied Systems) European Project, Contract Number FP6-511931

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3632: R. Nieuwenhuis (Ed.), Automated Deduction – CADE-20. XIII, 459 pages. 2005.

Vol. 3626: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis. X, 349 pages. 2005.

Vol. 3607: J.-D. Zucker, L. Saitta (Eds.), Abstraction, Reformulation and Approximation. XII, 376 pages. 2005.

Vol. 3596: F. Dau, M.-L. Mugnier, G. Stumme (Eds.), Conceptual Structures: Common Semantics for Sharing Knowledge. XI, 467 pages. 2005.

Vol. 3587: P. Perner, A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition. XVII, 695 pages. 2005.

Vol. 3584: X. Li, S. Wang, Z.Y. Dong (Eds.), Advanced Data Mining and Applications. XIX, 835 pages. 2005.

Vol. 3581: S. Miksch, J. Hunter, E. Keravnou (Eds.), Artificial Intelligence in Medicine. XVII, 547 pages. 2005.

Vol. 3577: R. Falcone, S. Barber, J. Sabater-Mir, M.P. Singh (Eds.), Trusting Agents for Trusting Electronic Societies. VIII, 235 pages. 2005.

Vol. 3575: S. Wermter, G. Palm, M. Elshaw (Eds.), Biomimetic Neural Learning for Intelligent Robots. IX, 383 pages. 2005.

Vol. 3571: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XVI, 1028 pages. 2005.

Vol. 3559: P. Auer, R. Meir (Eds.), Learning Theory. XI, 692 pages. 2005.

Vol. 3558: V. Torra, Y. Narukawa, S. Miyamoto (Eds.), Modeling Decisions for Artificial Intelligence. XII, 470 pages. 2005.

Vol. 3554: A. Dey, B. Kokinov, D. Leake, R. Turner (Eds.), Modeling and Using Context. XIV, 572 pages. 2005.

Vol. 3539: K. Morik, J.-F. Boulicaut, A. Siebes (Eds.), Local Pattern Detection. XI, 233 pages. 2005.

Vol. 3538: L. Ardissono, P. Brna, A. Mitrovic (Eds.), User Modeling 2005. XVI, 533 pages. 2005.

Vol. 3533: M. Ali, F. Esposito (Eds.), Innovations in Applied Artificial Intelligence. XX, 858 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005.

Vol. 3505: V. Gorodetsky, J. Liu, V. A. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005.

Vol. 3492: P. Blache, E. Stabler, J. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005.

Vol. 3476: J. Leite, A. Omicini, P. Torroni, P. Yolum (Eds.), Declarative Agent Languages and Technologies II. XII, 289 pages. 2005.

Vol. 3464: S.A. Brueckner, G.D.M. Serugendo, A. Karageorgos, R. Nagpal (Eds.), Engineering Self-Organising Systems. XIII, 299 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005.

Vol. 3451: M.-P. Gleizes, A. Omicini, F. Zambonelli (Eds.), Engineering Societies in the Agents World. XIII, 349 pages. 2005.

Vol. 3446: T. Ishida, L. Gasser, H. Nakashima (Eds.), Massively Multi-Agent Systems I. XI, 349 pages. 2005.

Vol. 3445: G. Chollet, A. Esposito, M. Faundez-Zanuy, M. Marinaro (Eds.), Nonlinear Speech Modeling and Applications. XIII, 433 pages. 2005.

Vol. 3438: H. Christiansen, P.R. Skadhauge, J. Villadsen (Eds.), Constraint Solving and Language Processing. VIII, 205 pages. 2005.

Vol. 3430: S. Tsumoto, T. Yamaguchi, M. Numao, H. Motoda (Eds.), Active Mining. XII, 349 pages. 2005.

Vol. 3419: B. Faltings, A. Petcu, F. Fages, F. Rossi (Eds.), Constraint Satisfaction and Constraint Logic Programming. X, 217 pages. 2005.

Vol. 3416: M. Böhlen, J. Gamper, W. Polasek, M.A. Wimmer (Eds.), E-Government: Towards Electronic Democracy. XIII, 311 pages. 2005.

Vol. 3415: P. Davidsson, B. Logan, K. Takadama (Eds.), Multi-Agent and Multi-Agent-Based Simulation. X, 265 pages. 2005.

Vol. 3403: B. Ganter, R. Godin (Eds.), Formal Concept Analysis. XI, 419 pages. 2005.

Vol. 3398: D.-K. Baik (Ed.), Systems Modeling and Simulation: Theory and Applications. XIV, 733 pages. 2005.

Vol. 3397: T.G. Kim (Ed.), Artificial Intelligence and Simulation. XV, 711 pages. 2005.

Vol. 3396: R.M. van Eijk, M.-P. Huget, F. Dignum (Eds.), Agent Communication. X, 261 pages. 2005.

Vol. 3394: D. Kudenko, D. Kazakov, E. Alonso (Eds.), Adaptive Agents and Multi-Agent Systems II. VIII, 313 pages. 2005.

Vol. 3392: D. Seipel, M. Hanus, U. Geske, O. Bartenstein (Eds.), Applications of Declarative Programming and Knowledge Management. X, 309 pages. 2005.

Vol. 3374: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems. X, 279 pages. 2005.

Vol. 3371: M.W. Barley, N. Kasabov (Eds.), Intelligent Agents and Multi-Agent Systems. X, 329 pages. 2005.

Vol. 3369: V. R. Benjamins, P. Casanovas, J. Breuker, A. Gangemi (Eds.), Law and the Semantic Web. XII, 249 pages. 2005.

Vol. 3366: I. Rahwan, P. Moraitis, C. Reed (Eds.), Argumentation in Multi-Agent Systems. XII, 263 pages. 2005.

Vol. 3359: G. Grieser, Y. Tanaka (Eds.), Intuitive Human Interfaces for Organizing and Accessing Intellectual Assets. XIV, 257 pages. 2005.

Vol. 3346: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 249 pages. 2005.

Vol. 3345: Y. Cai (Ed.), Ambient Intelligence for Scientific Discovery. XII, 311 pages. 2005.

Vol. 3343: C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, T. Barkowsky (Eds.), Spatial Cognition IV. XIII, 519 pages. 2005.

Vol. 3339: G.I. Webb, X. Yu (Eds.), AI 2004: Advances in Artificial Intelligence. XXII, 1272 pages. 2004.

Vol. 3336: D. Karagiannis, U. Reimer (Eds.), Practical Aspects of Knowledge Management. X, 523 pages. 2004.

Vol. 3327: Y. Shi, W. Xu, Z. Chen (Eds.), Data Mining and Knowledge Management. XIII, 263 pages. 2005.

Vol. 3315: C. Lemaître, C.A. Reyes, J.A. González (Eds.), Advances in Artificial Intelligence – IBERAMIA 2004. XX, 987 pages. 2004.

Vol. 3303: J.A. López, E. Benfenati, W. Dubitzky (Eds.), Knowledge Exploration in Life Science Informatics. X, 249 pages. 2004.

Vol. 3301: G. Kern-Isberner, W. Rödder, F. Kulmann (Eds.), Conditionals, Information, and Inference. XII, 219 pages. 2005.

Vol. 3276: D. Nardi, M. Riedmiller, C. Sammut, J. Santos-Victor (Eds.), RoboCup 2004: Robot Soccer World Cup VIII. XVIII, 678 pages. 2005.

Vol. 3275: P. Perner (Ed.), Advances in Data Mining. VIII, 173 pages. 2004.

Vol. 3265: R.E. Frederking, K.B. Taylor (Eds.), Machine Translation: From Real Users to Research. XI, 392 pages. 2004.

Vol. 3264: G. Paliouras, Y. Sakakibara (Eds.), Grammatical Inference: Algorithms and Applications. XI, 291 pages. 2004.

Vol. 3259: J. Dix, J. Leite (Eds.), Computational Logic in Multi-Agent Systems. XII, 251 pages. 2004.

Vol. 3257: E. Motta, N.R. Shadbolt, A. Stutt, N. Gibbins (Eds.), Engineering Knowledge in the Age of the Semantic Web. XVII, 517 pages. 2004.

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3248: K.-Y. Su, J. Tsujii, J.-H. Lee, O.Y. Kwong (Eds.), Natural Language Processing – IJCNLP 2004. XVIII, 817 pages. 2005.

Vol. 3245: E. Suzuki, S. Arikawa (Eds.), Discovery Science. XIV, 430 pages. 2004.

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3230: J.L. Vicedo, P. Martínez-Barco, R. Muñoz, M. Saiz Noeda (Eds.), Advances in Natural Language Processing. XII, 488 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3228: M.G. Hinchey, J.L. Rash, W.F. Truszkowski, C.A. Rouff (Eds.), Formal Approaches to Agent-Based Systems. VIII, 290 pages. 2004.

Vol. 3215: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. LVII, 906 pages. 2004.

Vol. 3214: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. LVIII, 1302 pages. 2004.

Vol. 3213: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LVIII, 1280 pages. 2004.

Vol. 3209: B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, G. Stumme (Eds.), Web Mining: From Web to Semantic Web. IX, 201 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3171: A.L. C. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004. XVII, 548 pages. 2004.

Table of Contents

Normative Multiagent Systems and Trust Dynamics

Guido Boella[1] and Leendert van der Torre[2]

[1] Dipartimento di Informatica. Università di Torino - Italy
guido@di.unito.it
[2] CWI Amsterdam and Delft University of Technology
torre@cwi.nl

Abstract. In this paper we use recursive modelling to formalize sanction-based obligations in a qualitative game theory. In particular, we formalize an agent who attributes mental attitudes such as goals and desires to the normative system which creates and enforces its obligations. The wishes (goals) of the normative system are the commands (obligations) of the agent. Since the agent is able to reason about the normative system's behavior, our model accounts for many ways in which an agent can violate a norm believing that it will not be sanctioned. We thus propose a cognitive theory of normative reasoning which can be applied in theories requiring dynamic trust to understand when it is necessary to revise it.

1 Introduction

Recently there has been interest in extending multiagent systems with concepts traditionally studied in deontic logic, such as obligations, permissions, rights, commitments, *et cetera*. In this paper we discuss the impact on trust of the theory behind our approach of obligations in virtual communities [1,2], which is based on two assumptions:

1. We define a theory of rational decision making in normative multiagent systems as a combination of multiagent systems and normative systems, for which we use recursive modelling and the attribution of mental attitudes to normative systems [2].
2. The role of deontic logic in our normative multiagent systems is to define the logic of the mental attitudes of the agents, for which we use input/output logics [3].

We focus on the motivations of agents when they violate norms. In particular, a sophisticated theory of trust dynamics would not increase trust if an agent only fulfills promises out of selfishness, because in the future the promise may not serve its needs. Analogously, it would not decrease the trust if the other agent did its best to fulfill the promise, but failed due to circumstances beyond its control. As applications of normative multiagent systems get more sophisticated, we need a more detailed model of rational decision making agents in such systems.

In Section 2 we discuss rational decision making, and in Section 3 the effect of violations on trust. In Section 4 we introduce the multiagent system, in Section 5 we define obligations and in Section 6 decisions of the agents, illustrated in Section 7.

R. Falcone et al. (Eds.): Trusting Agents, LNAI 3577, pp. 1–17, 2005.

2 Rational Decision Making in Normative Multiagent Systems

A theory of rational decision making is essential for many theories and applications in which agents are able to violate norms and such norm violations have consequences, such as theories of trust and reputation, fraud and deception [4], threats in persuasion [5] electronic commerce, virtual communities [1,6], social agents [7], agent-based software engineering [8,9], *et cetera*. This can be opposed to the Shoham and Tennenholtz' characterization of social laws in game theory [10], because their theory may be useful to study the use of norms in games and the emergence of norms in multiagent systems, but it is less useful to study the effectiveness of norms. We thus assume that norms are represented as soft constraints, which are used in detective control systems where violations can be detected (you can enter a train without a ticket, but you may be checked and sanctioned), instead of hard constraints, which are restricted to preventative control systems that are built such that violations are impossible (you cannot enter a metro station without a ticket).

Normally an agent fulfills its obligations, because otherwise its behavior counts as a violation that is being sanctioned, and the agent dislikes sanctions. Moreover, it may not like that its behavior counts as a violation regardless of the sanction, and it may act according to the norm regardless of whether its behavior counts as a violation, because it believes that this is fruitful for itself or for the community it belongs to. There are three categories of exceptions to this normal behavior.

First, an agent may violate an obligation when the violation is preferred to the sanction, like people who do not care about speeding tickets. In such cases of norm violations, the system may wish to increase the sanctions associated with the norms which are violated (for speeding) or decrease the sanction of another norm (death penalty).

Secondly, the agent may have conflicting desires, goals or obligations which it considers more important. Obligations may conflict with the agent's private preferences like wishes, desires, and goals. In the case of conflict with goals the agent has committed to, the agent has to decide whether the obligation is preferred to the goal, as well as whether it should change its mind. Moreover, even respectful agents, that always first try to fulfill their obligations before they consider their private preferences, do not fulfill all obligations in the case of contradictory obligations.

Thirdly, the agent may think that its behavior does not count as a violation, or that it will not be sanctioned. This may be the case when the normative system has no advantage in sanctioning the agent. One possibility is that the sanction has a cost for the system which overcomes the damage done by the violation: e.g., the sanction for not having payed taxes may be applied only above a certain threshold of money. But more likely the absence of violation is due to an action of the agent. The agent can change the system's behavior by affecting its desires or its goals. Its action may abort the goal of the normative system to count the agent's behavior as a violation or to sanction it, as in the case of bribing, or it may trigger a conflict for the normative system. The agent can use recursive modelling to exploit desires and goals of the normative system thus modifying its motivations and inducing it not to decide for the sanction.

3 Trust

There are many definitions of trust: the agents have a goal which can be achieved by means of the other agent's action [4]. They "bet" on the behavior of the other agents [11], since they could get the same good in other ways (i.e., buying the same good at an higher price from another agent they know already) [12,13]. One prominent class of trust scenarios is when an agent believes that the trustee "is under an obligation to do Z". In these situations, according to Jones [14] "trust amounts to belief in de facto conformity to normative requirements". Jones [14] aims at providing an "identifiable core" of this concept. He builds a classification of scenarios involving trust based on two parameters: the "rule belief" - the belief that exists a regularity in the trustee's behavior - and the "conformity belief" - the belief that "this regularity will again be instantiated on some given occasion".

Less attention, instead, has been devoted to trust dynamics. In particular, Falcone and Castelfranchi [15] notice that many approaches to trust dynamics adopt a naïve view where:

> "to experiences to each success of the trustee corresponds an increment in the amount of the trustier's trust towards it, and vice versa, to every trustee's failure corresponds a reduction of the trustiers trust towards the trustee itself.

They argue that the motivation of this simplification rests in the lack of cognitive models of trust:

> "this primitive view cannot be avoided till Trust is modelled just as a simple index, a dimension, a number; for example, reduced to mere subjective probability. We claim that a cognitive attribution process is needed in order to update trust on the basis of an 'interpretation of the outcome of A's reliance on B and of B's performance. [...] In particular we claim that the effect of both B's failure or success on A's Trust in B depends on A's 'causal attribution' of the event. Following 'causal attribution theory' any success or failure can be either ascribed to factors internal to the subject, or to environmental, external causes, and either to occasional facts, or to stable properties."

The cognitive model of trust of [15] is based on a portrait of the mental state of trust in cognitive terms (beliefs, goals). Their model includes two main basic beliefs. First, a competence belief which includes a sufficient evaluation of Y's abilities is, that X should believe that Y is useful for this goal of its, that Y can produce/provide the expected result, and that Y can play such a role in X's plan/action. Second a willingness belief where X should think that Y not only is able and can do that action/task, but Y actually will do what X needs This belief makes the trustee's behavior predictable and includes the trustees reasons and motives for complying. In particular, X believes that Y has some motives for helping it (for adopting its goal), and that these motives will probably prevail -in case of conflict- on other motives. Notice that motives inducing adoption are of several different kinds: from friendship to altruism, from morality to fear of sanctions, from exchange to cooperation. Moreover, when X trusts someone, X

is in a strategic situation: X believes that there is interference and that his rewards, the results of his projects, depend on the actions of another agent Y.

Since we propose a cognitive theory of decisions under norms, which, as said above, constitute one prominent case of trust situations, our theory can be useful for a dynamic theory of trust. In particular, we can consider two respects:

- A structural dimension for characterizing trust and its dynamics.
- A behavior dimension relating recursive modelling and trust dynamics.

First, a characterization of trust dynamics should be based on norm behavior: trust should be increased if an agent followed its commitments and obligations, while it should be decreased in case norms are violated. But as we discussed, norms can be violated for different reasons, which should be considered in the dynamic adjustment of trust:

- Behavioristic trust dynamics: increase trust when the norm is fulfilled, decrease when the norm is violated.
- Sanction trust dynamics: decrease trust if the norm is violated if the associated sanction does not provide a motivation for the trustee.
- Goal trust dynamics: increase trust when the agent has a goal to fulfil the norm, even if it is not able to respect the norm for some external reason, decrease when it has no goal to fulfil norm or the goal is conflicting with more important goals.
- Desire trust dynamics: increase trust when the agent desires to fulfill the obligations; desires represent the inner motivations of an agent, while goals can be adopted from other agents' ones.
- Cooperative trust dynamics: in case of norm violation the agent informs the other agents (an example of mutual support or mutual responsiveness). This problem is addressed, e.g., [16,17,18].

The second dimension concerns the behavior of agents. If the fact that the agents have a goal which can be achieved by means of the other agent's action is at the basis of trust, a trust situation is inherently a strategic situation, as highlighted by [4,15]. Two agents are in a strategic situation if an agent "believes that there is an interference and that her rewards, the results of her projects, depend on the actions of another agent". For this reason, an agent must have a profile of the other agents (apart from the first case above, where it can simply observe its behavior). As we discuss in Section 6, a basic ability of agents is to recursively model the behavior of the other agents. In our model this is at the basis of the definition of obligation: only by recursively modelling the decision of the normative system an agent can understand whether it will be sanctioned or not. Thus, to build a dynamic model of trust under obligations, it is useful to have a complete model of the decision making under obligations which includes also the recursive modelling abilities of agents.

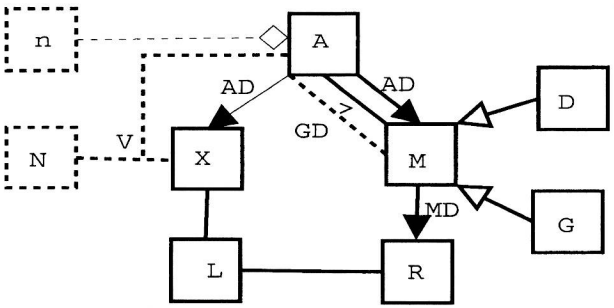


Fig. 1. Conceptual model of normative multiagent system

4 Normative Multiagent System

The conceptual model of the normative multiagent system is visualized in Figure 1. Following the usual conventions of for example class diagrams in the unified modelling language (UML), □ is a concept or set, — and → are associations between concepts, —▷ is the "is-a" or subset relation, and —◇ is a relation called "part-of" or aggregation. The logical structure of the associations is detailed in the definitions below.

We first explain the multiagent system and thereafter the normative extension. Agents (A) are described (AD) by actions called *decision variables* (X) and by motivations (M) guiding its decision making. The motivational state of an agent is composed by its desires (D) and goals (G). Agents may share decision variables, desires or goals, though this is not used in the games discussed in this paper. Desire and goal rules can be conflicting, and the way the agent resolves its conflicts is described by a priority relation ($\geq$) that expresses its agent characteristics [19]. The priority relation is defined on the powerset of the motivations such that a wide range of characteristics can be described, including social agents that take the desires or goals of other agents into account. The priority relation contains at least the subset-relation, which expresses a kind of independence between the motivations.

Definition 1 (Agent set). *An agent set is a tuple* $\langle A, X, D, G, AD, \geq \rangle$*, where*

- *the agents* A*, decision variables* X*, desires* D *and goals* G *are four finite disjoint sets.* $M = D \cup G$ *are the motivations defined as the union of the desires and goals.*
- *an agent description* $AD : A \rightarrow 2^{X \cup D \cup G}$ *is a complete function that maps each agent to sets of decision variables, desires and goals, such that each decision variable is assigned to at least one agent. For each agent* $a \in A$*, we write* X_a *for* $X \cap AD(a)$*,* D_a *for* $D \cap AD(a)$*,* G_a *for* $G \cap AD(a)$*.*
- *a priority relation* $\geq$*:* $A \rightarrow 2^M \times 2^M$ *is a function from agents to a transitive and reflexive relation on the powerset of the motivations containing at least the subset relation. We write* $\geq_a$ *for* $\geq (a)$*.*

Desires and goals are abstract concepts which are described (MD) by – though conceptually not identified with – rules (R) built from literals (L). Rules consist of

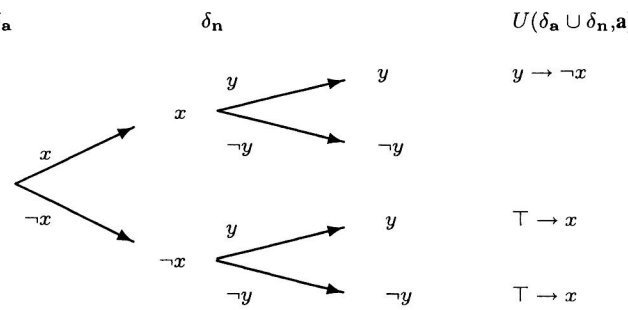


Fig. 2. The game between agent **a** and agent **n**

an antecedent (body, input) and a consequent (head, output), which are in our case respectively a set of literals and a literal. This simple structure of rules keeps the formal exposition simple and is sufficient for our purposes here, for the extension of rules to pairs of propositional sentences we can use input/output logics [3]. As priorities are associated with mental attitudes instead of rules, the priority of a rule associated with a desire may be different from the priority of the same rule associated with a goal. We do not use more complex constructions of rules, as used for example in logic programming, nonmonotonic reasoning or in description logics, because we do not seem to need the additional complexity and moreover, these rules typically focus on a limited set of reasoning patterns which cannot be used for our purposes. In particular, they assume the identity rule 'if p then p', see [3,20] for a discussion. It is well known that desires are different from goals, and we can adopt distinct logical properties for them. For example, goals can be adopted from other agents, whereas desires cannot. In this paper we do not make any additional assumptions on desires and goals, and we thus do not formally characterize the distinction between desires and goals, because it is beyond the scope of this paper.

Definition 2 (MAS). *A multiagent system MAS is a tuple $\langle A, X, D, G, AD, MD, \geq \rangle$:*

- *the set of literals built from X, written as $L(X)$, is $X \cup \{\neg x \mid x \in X\}$, and the set of rules built from X, written as $R(X) = 2^{L(X)} \times L(X)$, is the set of pairs of a set of literals built from X and a literal built from X, written as $\{l_1, \ldots, l_n\} \rightarrow l$. We also write $l_1 \wedge \ldots \wedge l_n \rightarrow l$ and when $n = 0$ we write $\top \rightarrow l$. Moreover, for $x \in X$ we write $\sim x$ for $\neg x$ and $\sim(\neg x)$ for x.*
- *the motivational description $MD : M \rightarrow R(X)$ is a complete function from the sets of desires and goals to the set of rules built from X. For a set of motivations $S \subseteq M$, we write $MD(S) = \{MD(s) \mid s \in S\}$.*

We illustrate the notation by the example visualized in Figure 2. In the example, there are two agents, called agent **a** and agent **n**, who in turn make a decision: agent **a** chooses between x and $\neg x$, and agent **n** chooses between y and $\neg y$.

Agent **a** desires x only when agent **n** does not do y. In the example, we only formalize the mental attitudes; how the agents make decisions and the meaning of the function U at the right hand side of Figure 2 are explained in Section 6.

Example 1. Consider a multiagent system $\langle A, X, D, G, AD, MD, \geq \rangle$ with $A = \{\mathbf{a}, \mathbf{n}\}$, $X_{\mathbf{a}} = \{x\}$, $X_{\mathbf{n}} = \{y\}$, $D_{\mathbf{a}} = \{d_1, d_2\}$, $D_{\mathbf{n}} = \{d_3\}$, $G = \emptyset$, $MD(d_1) = \top \rightarrow x$, $MD(d_2) = y \rightarrow \neg x$, $MD(d_3) = \top \rightarrow y$, $\geq_a$ is such that $\{\top \rightarrow x, y \rightarrow \neg x\} \geq \{y \rightarrow \neg x\} \geq \{\top \rightarrow x\} \geq \emptyset$. Agent $\mathbf{a}$ could also consider d_3 in his priority ordering, but we assume that d_3 does not have any impact on it. Agent a desires unconditionally to decide to do x, but if agent $\mathbf{n}$ decides y, then agent $\mathbf{a}$ desires $\neg x$, since the second rule is preferred over the first one in the ordering. Agent $\mathbf{n}$ desires to do y.

To describe the normative system, we introduce several additional items. The basic idea of the formalization of our normative multiagent system is that the normative system can be modelled as an agent, and therefore mental attitudes like desires and goals can be attributed to it, because it is autonomous and it possesses several other properties typically attributed to agents. Example 1 can be interpreted as a game between an agent and its normative system instead of between two ordinary agents. In the context of this paper, this idea is treated as a useful way to combine multiagent systems and normative systems, though it can also be defended from a more philosophical point of view. A motivation has been discussed in [21], though in that paper we did not give a formalization of normative multiagent systems as we do in this paper. First we identify one agent in the set of agents as the normative agent. We represent this normative agent by $\mathbf{n} \in A$. Moreover, we introduce a set of norms N and a norm description V that associates violations with decision variables of the normative agent. Finally, we associate with each agent some of the goals GD of the normative agent, which represents the goals this agent is considered responsible for. Note that several agents can be responsible for the same goal, and that there can be goals no agent is considered responsible for. We do not assume that agents can only be responsible for their own decisions. In some more complex social phenomena agents may also be responsible for other agents' decisions, and this assumption may be relaxed in the obvious way. For example, in some legal systems, parents are responsible for actions concerning their children, or the owners of artificial agents are responsible for the actions the agents perform on their behalf.

Definition 3 (Norm description). *A normative multiagent system $NMAS$ is a tuple $\langle A, X, D, G, AD, MD, \geq, \mathbf{n}, N, V, GD \rangle$, where $\langle A, X, D, G, AD, MD, \geq \rangle$ is a multiagent system, and:*

- *the normative agent $\mathbf{n} \in A$ is an agent.*
- *the norms $\{n_1, \ldots, n_m\} = N$ is a set disjoint from A, X, D, and G.*
- *the norm description $V : N \times A \rightarrow X_{\mathbf{n}}$ is a complete function from the norms to the decision variables of the normative agent: we write $V(n, a)$ for the decision variable which represents that there is a violation of norm n by agent $a \in A$.*
- *the goal distribution $GD : A \rightarrow 2^{G_{\mathbf{n}}}$ is a function from the agents to the powerset of the goals of the normative agent, where $GD(a) \subseteq G_{\mathbf{n}}$ represents the goals of agent $\mathbf{n}$ the agent a is responsible for.*